DOCUMENT RESUME

ED 386 476 TM 024 030

AUTHOR Zwick, Rebecca
TITLE Pairwise Comparison Procedures for One-Way Analysis of Variance Designs. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-91-30
PUB DATE Apr 91
CONTRACT 5-T32-MH15745
NOTE 62p.; Revised version of an article "Testing Pairwise Contrasts in One-Way Analysis of Variance Designs" in "Psychoneuroendocrinology," v11, p253-76, 1986.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Comparative Analysis; Factor Structure; *Research Design
IDENTIFIERS Power (Statistics); Type I Errors

ABSTRACT

Research in the behavioral and health sciences frequently involves the application of one-factor analysis of variance models. The goal may be to compare several independent groups of subjects on a quantitative dependent variable or to compare measurements made on a single group of subjects on different occasions or under different conditions. In analyzing data of this kind, it is usually of interest to determine which pairs of population means are likely to differ. In this paper, the selection of pairwise multiple comparison procedures for one-way analysis of variance designs is considered, following a discussion of Type I error and power issues as they apply to the testing of multiple hypotheses. Procedures are included which are appropriate when normality or variance homogeneity assumptions are violated. The focus is on procedures that are easy to understand and apply. Single-step procedures are emphasized because of their simplicity and because they allow for the construction of confidence intervals. (Contains 5 tables and 69 references.) (Author)

RESEARCH REPORT

PAIRWISE COMPARISON PROCEDURES FOR ONE-WAY ANALYSIS OF VARIANCE DESIGNS

Rebecca Zwick

Educational Testing Service
Princeton, New Jersey
April 1991

Pairwise Comparison Procedures for One-Way Analysis of Variance Designs

Rebecca Zwick

Educational Testing Service

To appear in Methodological and Quantitative Issues in the Analysis of Psychological Data, edited by Gideon Keren and Charles Lewis.

This chapter is a revised version of an article, "Testing Pairwise Contrasts in One-Way Analysis of Variance Designs," that appeared in Psychoneuroendocrinology, Volume 11, pp. 253-276, 1986. It appears here with the permission of Pergamon Press. The initial research was supported in part by a National Research Service Award (No. 5-T32-MH15745) from the National Institute of Mental Health to the University of North Carolina at Chapel Hill. Preparation of the current version was supported in part by Educational Testing Service. Special thanks are due to Juliet Shaffer for her valuable and detailed recommendations for improving this chapter. In addition, I would like to thank C. Clifford Attkisson, Henry Braun, and Peter Vitaliano for their comments on the original article, Charlie Lewis and Jo-Ling Liang for comments on the current version, and Kay Tyberg for preparation of the manuscript.

Abstract

Research in the behavioral and health sciences frequently involves the application of one-factor analysis of variance models. The goal may be to compare several independent groups of subjects on a quantitative dependent variable or to compare measurements made on a single group of subjects on different occasions or under different conditions. In analyzing data of this kind, it is usually of interest to determine which pairs of population means are likely to differ. In this paper, the selection of pairwise multiple comparison procedures for one-way analysis of variance designs is considered, following a discussion of Type I error and power issues as they apply to the testing of multiple hypotheses. Procedures are included which are appropriate when normality or variance homogeneity assumptions are violated. The focus is on procedures that are easy to understand and apply. Single-step procedures are emphasized because of their simplicity and because they allow for the construction of confidence intervals.

Pairwise Comparison Procedures for One-Way Analysis of Variance Designs

Research in the behavioral and health sciences frequently involves the application of one-factor analysis of variance (ANOVA) models. The goal may be to compare several independent groups of subjects on a quantitative dependent variable or, alternatively, to compare measurements made on different occasions or under different conditions on a single group of subjects. If there is reason to believe that there are differences among the groups (or occasions or conditions), the researcher frequently wishes to compare the means in a pairwise fashion. Although the procedures for conducting omnibus hypothesis tests for one-factor ANOVA models are familiar to most researchers, the issues that must be considered in choosing pairwise multiple comparison procedures (MCPs) are not as well-understood. In this paper, the selection of pairwise MCPs for one-factor ANOVA models is considered, following a discussion of Type I error and power issues as they apply to the testing of multiple hypotheses. Although the paper focuses on the independent-sample case, repeated measures models are considered briefly as well.

Type I Error and Power

Any student who has taken an elementary statistics course can recite the definitions of Type I error and power: The Type I error rate is the

probability of rejecting the null hypothesis when the null hypothesis is true and power is the probability of rejecting the null hypothesis when the null hypothesis is false. However, these concepts become much more complex when applied to multiple hypothesis tests, such as MCPs. In the multiple-comparison case, it is possible to define many varieties of Type I error rates (Bernhardson, 1975; Zwick & Marascuilo, 1984). Two of the most important (defined here in terms of pairwise MCPs only) are the comparisonwise error rate, α_C, which is the probability of making a Type I error on a particular comparison, and the experimentwise error rate, α_E, which is the probability of making at least one Type I error in conducting the entire set of pairwise comparisons associated with an experiment. (For an experiment with k means, there are $k(k - 1)/2$ distinct pairwise comparisons.)

Some MCPs are designed to allow direct control of the comparisonwise error rate; that is, the researcher sets a nominal Type I error rate for each comparison. In other methods, the researcher determines a nominal value for the experimentwise error rate. If a method that allows direct control of α_C is chosen, probability inequalities such as the Bonferroni inequality can be used to calculate an upper bound for α_E. The Bonferroni inequality, as applied in this context, states that if $k(k - 1)/2$ pairwise comparisons are performed, each with a Type I error probability equal to α_C, then the experimentwise Type I error rate, α_E, will be less than or equal to $[k(k - 1)/2]\alpha_C$. That is, the experimentwise error rate is less than or equal to the sum of the comparisonwise error rates. (This upper bound can exceed 1, whereas α_E, of course, can not.) Other probability

inequalities, such as the Dunn-Šidák inequality (Dunn, 1958, 1959, 1974; Šidák, 1967) can be used to produce a more refined upper bound in certain cases. The Bonferroni inequality, however, has the advantage of simplicity and generality.

It is important to note that experimentwise and comparisonwise error rates are not simply interchangeable ways of evaluating Type I error. This can be illustrated with an example. Suppose we used computer simulation techniques to investigate the Type I error rates of two competing MCPs. In order to study Type I error in this way, random numbers are generated and assigned to groups. (Because there are no "population" differences among the groups, all statistically significant comparisons will be Type I errors.) The test statistics of interest are then performed on the random data. Suppose that the results for 100 simulated experiments with $k = 3$ groups and a nominal α_E of .05 are as shown in Table 1.

Insert Table 1 about here

We can calculate an empirical estimate ($\hat{\alpha}_E$) of the experimentwise error rate for each of the two MCPs as follows:

$$[1] \qquad \hat{\alpha}_E = \frac{\text{Number of experiments with at least one significant pairwise comparison}}{\text{Number of experiments}} .$$

In calculating $\hat{\alpha}_E$, the 100 experiments are divided into two classes: those that have no Type I errors and those that have one or more Type I errors. The value of $\hat{\alpha}_E$ is simply the proportion of experiments in the second class. For each of the MCPs, $\hat{\alpha}_E = .05$; that is, the estimated

experimentwise error rate is equal to the nominal α_E. We therefore expect that when the null hypothesis is true, application of either of these MCPs will lead to at least one Type I error in five percent of the experiments performed. The value of $\hat{\alpha}_E$ tells us nothing about the likelihood of a Type I error on a particular comparison. The estimated comparisonwise error rate for this example can be calculated as follows:

$$[2] \quad \hat{\alpha}_C = \frac{\text{Number of significant pairwise comparisons}}{[k(k-1)/2] \cdot [\text{Number of experiments}]}$$

Note that this is simply the overall proportion of pairwise comparisons that resulted in Type I errors. For MCP 1, $\hat{\alpha}_C = [1(4) + 2(1)]/3(100) = .02$; for MCP 2, $\hat{\alpha}_C = 3(5)/3(100) = .05$. Therefore, although the MCPs are identical in terms of $\hat{\alpha}_E$, they differ in terms of $\hat{\alpha}_C$. Which of these error rates is most useful to the applied researcher?

If, as in most cases, the research conclusions depend on the simultaneous correctness of the set of $k(k-1)/2$ inferences, experimentwise error control is appropriate; if the researcher is concerned instead about the correctness of individual inferences about pairs of means, an MCP that allows direct control of the comparisonwise error rate should be selected. An example of a case in which comparisonwise control might be preferable is as follows: Suppose a researcher conducts a study in which three groups, A, B, and C, are compared with $\alpha_E = .05$. The researcher then conducts a similar study which includes two additional groups, D and E, although her primary interest is still in groups A, B, and C. If she again uses $\alpha_E = .05$, her tests of the three pairwise differences

among groups A, B, and C will be more conservative (i.e., less likely to lead to statistically significant results) than in the previous study because the experimental error rate of .05 will be allocated among a larger number of comparisons. Therefore, it might be considered desirable to hold the value of α_C, rather than α_E, constant across studies. Even if this rationale were applied, however, it would still be important for the researcher to be aware of the experimentwise error rate. That is, if the researcher decides to set α_C equal to .02, she should be prepared to accept an experimentwise error rate as large as $[k(k - 1)/2](.02) = 10(.02) = .20$ for the five-group study, assuming all pairwise comparisons are to be conducted.

Just as there are several kinds of Type I error rates that are pertinent to the choice of MCPs, there are several definitions of power that may be useful as well. For instance, for a set of three means with population values 1, 2, and 10, we could consider the probability of detecting one or more of these differences (any-pair power) or the probability of detecting all three pairwise differences (all-pairs power). These definitions will not be explained in detail here; a good discussion is given by Ramsey (1981; see also Einot & Gabriel, 1975; Gabriel, 1978; Ramsey, 1978).

Comparing the Empirical Type I Error Rates and Powers of Competing MCPs

Simulation studies like the hypothetical one described above are often performed in order to compare empirical estimates of the true Type I error

rates and powers associated with competing MCPs. Unfortunately, many published studies are misleading because they are flawed in design or interpretation. For example, investigators conducting simulation studies of competing MCPs have often failed to distinguish between procedures that provide comparisonwise error control and those that control Type I errors in an experimentwise fashion. It is not unusual to find an MCP with a nominal α_C of .05 being compared to a procedure with a nominal α_E of .05 (Einot & Gabriel, 1975; Zwick & Marascuilo, 1984). Comparisons of this kind provide no useful information about the relative performance of the MCPs. Even without performing a simulation study, it can be predicted that a procedure with a nominal α_C of .05 will produce more Type I errors than a procedure with a nominal α_E of .05. To achieve a more useful comparison of procedures that provide experimentwise control with those that provide comparisonwise control, MCPs with a nominal α_E of .05 should be compared with MCPs with a nominal α_C of $.05/[k(k - 1)/2]$. The MCPs that controls α_C can then be regarded as having a nominal α_E of approximately .05.[1]

Similarly, power comparisons can be meaningfully interpreted only if the MCPs under evaluation have the same nominal α_E. This is because the probability of rejecting a false null hypothesis (power) can always be made larger by increasing α. Thus, a procedure with a nominal α_C of .05 is expected to lead to more rejections of false null hypotheses than a procedure with a nominal α_E of .05 because the former procedure is known to have a larger experimentwise error rate. This does not mean that the former procedure is more powerful in any practical sense. (If meaningful power increases could be achieved by increasing the Type I error rate, we

could simply set the nominal α equal to 1.00. By always rejecting the null hypothesis, we would be assured of detecting any true differences!) A related point that is often overlooked is that in making power comparisons, it is important to consider whether the true Type I error rates for the procedures being compared depart substantially from the nominal α. That is, even if the nominal α_E is equal to .05 for two MCPs, it may be that the true error rate for one of the procedures is known to exceed the nominal α_E (as is the case with the protected t-test procedure, discussed below), whereas the other MCP does, in fact, control the error rate at the nominal α_E. Here again, it would be a mistake to conclude that the former procedure was more powerful.

These points have important implications for the MCP user, who may be tempted to pick the MCP that tends to yield the largest number of statistically significant differences without determining whether the apparent power superiority is, in fact, achieved at the expense of an increased risk of Type I error.

The Independent-Sample Case

One of the most common ANOVA designs involves a comparison of k independent groups of subjects on a quantitative dependent variable. In this section, five of the most common MCPs that are applicable to this model are described and illustrated using a hypothetical example. Like the ANOVA F test, these methods require the assumption that observations are independent random samples from normal populations with equal variances.

The inclusion of an MCP in the discussion below does not constitute a recommendation. Some of the MCPs were selected to be representative of certain types of procedures or of particular philosophies of Type I error control. A detailed evaluation of these five methods is provided, followed by a discussion of MCPs for use when normality or variance equality are thought to be violated.

Example

Suppose we are interested in comparing different forms of psychiatric treatment for psychotic inpatients. We choose a random sample of 40 psychotics and then randomly assign them to one of four forms of treatment: pharmacologic therapy (P), which involves the administration of anti-psychotic drugs; group psychotherapy (G); individual psychotherapy (I); and a combination of anti-psychotic drugs and individual psychotherapy (C). After one month, we ask an experienced clinical researcher to rate each patient on a series of items pertaining to the patient's ability to perform everyday tasks, maintain personal relationships, and hold a job. The rating scale yields overall scores ranging from 0 to 100, with higher scores indicating greater ability to function. The scores, means, variances, and sample sizes are as shown in Table 2.

Insert Table 2 about here

These hypothetical data will be used to illustrate the multiple comparison methods presented in this paper.

MCPs for the Normal Equal-Variance Case

In this section, the following MCPs are discussed: a) Scheffé's (1953) procedure, b) Tukey's (1953) Studentized range test and the Tukey-Kramer (Kramer, 1956, Tukey, 1953) modification for unequal sample sizes, c) the Dunn-Bonferroni method (Dunn, 1961), d) Fisher's (1935) protected t-test procedure, and e) the Newman-Keuls (Keuls, 1952; Newman, 1939) test. All five of these procedures are described in Hochberg and Tamhane (1987), Kirk (1982), and Miller (1981). It is important to note that these five methods do not differ in terms of the formulation of the test statistics used to make pairwise comparisons. In each case the test statistic can be written as follows:

[3] $$t_{ii'} = \frac{\bar{X}_i - \bar{X}_{i'}}{\sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}},$$

where $\bar{X}_i$ and $\bar{X}_{i'}$ are the two means being compared, n_i and $n_{i'}$ are the sample sizes associated with these two means, and MSW is the mean square within groups for the entire k-sample study, defined as $\sum_{i=1}^{k} (n_i - 1)s_i^2/(N - k)$ where k is the number of groups, $N = \sum_{i=1}^{k} n_i$ is the overall sample size, and s_i^2 is the variance in each sample. Because the example involves samples of equal size, Equation 3 can be simplified as follows:

[4] $$t_{ll'} = \frac{\bar{X}_l - \bar{X}_{l'}}{\sqrt{MSW\left(\frac{2}{n}\right)}},$$

where n = 10 is the sample size in each of the k = 4 groups. For instance, the test statistic for comparing group therapy (G) to the combination of pharmacologic and individual therapy (C) is:

[5] $$\frac{\bar{X}_G - \bar{X}_C}{\sqrt{MSW\left(\frac{2}{n}\right)}} = \frac{25 - 41}{\sqrt{96.83\left(\frac{2}{10}\right)}} = -3.64$$

The values of $t_{ll'}$ for the remaining five pairwise comparisons are given in Table 2.

The difference among the five MCPs listed above lies in a) the rules, if any, used to determine whether a given comparison is to be performed and b) the choice of critical values to which the t-statistics are to be compared. These aspects of each of the five procedures are described below.

Scheffé's Procedure. Scheffé comparisons are ordinarily performed after a significant ANOVA F-test. That is, the statistic F = MSB/MSW is first computed, where the mean square between groups MSB, is equal to $\sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$, $\bar{X}$ is the grand mean, and MSW is defined as above. Then if the observed value of F exceeds $F_{k-1,N-k;1-\alpha_E}$ the critical value of F with k - 1 and N - k degrees of freedom for the desired α_E level, pairwise t-tests of

the form shown in Equation 3 (or Equation 4) are performed, using as a critical value

[6] $$S_{k-1,N-k;1-\alpha_E} = \sqrt{(k-1)\,F_{k-1,N-k;1-\alpha_E}}$$

The comparison is significant if $|t| > S$ (i.e., if $t < -S$ or $t > S$). The Scheffé method controls the experimentwise Type I error rate at α_E regardless of the number of pairwise comparisons performed. In fact, the method controls the experimentwise error rate at α_E for the set of all contrasts. A contrast is a linear combination of population means (μ_i) of the form $a_1\mu_1 + a_2\mu_2 + \ldots + a_k\mu_k$, where the a_i are weights chosen so that $\sum_{i=1}^{k} a_i = 0$. For example, we might want to test the hypothesis that $\mu_1 - \frac{\mu_2 + \mu_3}{2} = 0$; that is, the mean of group 1 is equal to the mean of groups 2 and 3 combined. This contrast would be estimated in the sample as $\bar{X}_1 - \frac{\bar{X}_2 + \bar{X}_3}{2} = (1)(\bar{X}_1) + (-1/2)(\bar{X}_2) + (-1/2)(\bar{X}_3)$. Here, $a_1 = 1$, $a_2 = -1/2$, and $a_3 = -1/2$. A pairwise comparison is a special case of a contrast, where the weights are $a_i = 1$ and $a_{i'} = -1$. Contrasts other than pairwise comparisons are called complex contrasts. Although the Scheffé method applies to all contrasts, it is often used even when only pairwise comparisons are of interest. It is important to understand the relationship between the overall F-test and the pairwise comparisons performed via the Scheffé method. Although a significant F ratio implies the existence of at least one significant contrast, it does not imply the existence of a significant pairwise comparison. Therefore, the finding of

a significant F-test, but no significant pairwise t-tests is not inconsistent with theory. It should also be noted that no additional risk of Type I error is incurred if the prior F-test is omitted. The F-test can, however, be useful as a labor-saving device because, if the F-test is not significant, no pairwise comparison (or other contrast) will be found significant using S as a critical value.

In the present example MSB = 536.67, the observed value of F is 5.54, and, for α_E = .05, $F_{k-1,N-k;1-\alpha_E} = F_{3,36;.95} = 2.87$.[2] Therefore, the null hypothesis of no group differences is rejected. Comparisons of the form shown in Equation 4 can be performed, with $S_{k-1,N-k;1-\alpha_E} = \sqrt{3(2.87)} = 2.93$ as a critical value. The only statistically significant pairwise comparisons are those between the P and G groups and between the C and G groups. We would therefore conclude that, although pharmacologic therapy or a combination of pharmacologic and individual therapy differ from group psychotherapy in terms of their impact on the functioning of psychotic patients, no other distinctions can be made among the various modes of therapy.

Examination of the means for the G and C groups indicates that there is a 16-point difference in favor of the combined therapy group. Instead of merely concluding that these groups differ, we may wish to make an inference about the size of the difference between the G and C group means in the population. We can do this by constructing $100(1 - \alpha_E)\%$ simultaneous confidence intervals of the following form:

$$[7] \quad (\bar{X}_i - \bar{X}_{i'}) - S_{k-1,N-k;1-\alpha_E}\sqrt{MSW\left(\frac{2}{n}\right)} < (\mu_i - \mu_{i'}) < (\bar{X}_i - \bar{X}_{i'}) + S_{k-1,N-k;1-\alpha_E}\sqrt{MSW\left(\frac{2}{n}\right)} .$$

Substituting in the values from the example, we can write the 95% confidence interval for $\mu_G - \mu_C$ as follows:

$$(25 - 41) - 2.93\sqrt{96.83\left(\frac{2}{10}\right)} < (\mu_G - \mu_C) < (25 - 41) + 2.93\sqrt{96.83\left(\frac{2}{10}\right)}$$

$$-28.89 < (\mu_G - \mu_C) < -3.11$$

Thus, we can state with 95% confidence, that in the population, the number of points by which the mean for the combined therapy group exceeds the mean for the group therapy group is between about 3 and 29. The reason that the Scheffé intervals are called <u>simultaneous</u> confidence intervals is related to the type of error control that characterizes the Scheffé procedure. As stated above, the probability of at least one Type I error is controlled at α_E. This implies that, before the experiment is conducted, the probability of no Type I errors is $1 - \alpha_E$. Therefore, after performing the experiment, we can state with $100(1 - \alpha_E)$% confidence that <u>all</u> statements of the form shown in Equation 7 are true. In fact, the 95% confidence statement applies to all contrasts, not merely pairwise comparisons. In interpreting the results of a study, confidence intervals are usually more valuable than hypothesis tests alone. They can help the researcher to determine whether results that are statistically significant have any practical importance. For instance, in the present example, the researcher must take into

consideration that the difference between the G and C groups may be as small as about 3 points, a quantity that may be insignificant from a clinical standpoint.

Tukey's Studentized Range Test and the Tukey-Kramer Modification. Tukey's Studentized range test, also called the Honestly Significant Difference (HSD) test or Wholly Significant Difference (WSD) test,[3] is usually described as follows: Find the largest and smallest sample means, compute

[8] $$T = \frac{\bar{X}_{max} - \bar{X}_{min}}{\sqrt{\frac{MSW}{n}}}$$

and compare T to a critical value, denoted as $q_{k,N-k;1-\alpha_E}$ based on the distribution of the Studentized range. If this value is statistically significant (i.e. $|T| > q$), perform all other pairwise comparisons in the same fashion. The test statistic in Equation 8 differs from that shown in Equation 4 by a factor of $\sqrt{2}$ in the denominator. Therefore, comparing T to $q_{k,N-k;1-\alpha_E}$ is the same as comparing the largest value of t_{ij} to $q_{k,N-k;1-\alpha_E}/\sqrt{2}$. In order to find the appropriate critical value for the example, we must enter a table of the percentiles of the Studentized range (e.g., see Kirk, 1982 or Miller, 1981) and find the critical value corresponding to the number of means in the overall experiment (k), the error degrees of freedom (N - k), and the desired level of α_E. For four means, 36 degrees of freedom, and $\alpha_E = .05$, the critical value of $q_{k,N-k;1-\alpha_E}$ is found by linear

interpolation to be approximately 3.81. Therefore, if we want to use test statistics of the form shown in Equation 4, our critical value is $3.81/\sqrt{2} = 2.70$. Because the largest value of $t_{H'}$, shown in Equation 5, exceeds the critical value, we conclude, as before, that combined and group therapy produce different results. Proceeding to the remaining five comparisons, we find that, as in the Scheffé procedure, the only other statistically significant comparison is that between the P and G groups.

Like the Scheffé MCP (see Scheffé, 1953), the Tukey procedure controls the experimentwise error rate at a nominal value of α_E for the set of all contrasts (Hochberg & Tamhane, 1987), which, of course, includes all pairwise comparisons. (Although complex contrasts can be performed via the Tukey approach, however, this is rarely done because the Bonferroni and Scheffé methods are usually more powerful for this kind of test [Miller, 1981]. Because a single procedure should be selected for all comparisons of interest, a researcher who wanted to test a substantial number of complex contrasts would be wise to select the Scheffé method.)

For the Tukey MCP, the "prior" test shown in Equation 8 is simply an evaluation of the largest pairwise difference. There is no theoretical reason that this test need be performed before the other pairwise comparisons. It can save computational labor, however, because, if this comparison is not significant, no other pairwise comparisons will be found significant. For the significant comparisons, simultaneous confidence intervals of the form shown in Equation 7 can be constructed, with $q_{k,N-k;1-\alpha_E}/\sqrt{2}$ replacing $S_{k-1,N-k;1-\alpha_E}$.

Although the HSD method per se is applicable in the case of equal sample sizes only, the Tukey-Kramer modification for unequal sample sizes has been shown to have an experimentwise Type I error rate that does not exceed the nominal α (Hayter, 1984; see Hochberg & Tamhane, 1987, pp. 91 - 93). Applying the Tukey-Kramer method is equivalent to comparing t-statistics of the form shown in Equation 3 to $q_{k,N-k;1-\alpha_E}/\sqrt{2}$. Note that the substitution of the harmonic mean of the sample sizes of all k groups for the n in the denominator of T in Equation 8, as recommended by Winer (1962), leads to a test with poor Type I error control (Hochberg & Tamhane, 1987).

Dunn-Bonferroni Method. Dunn (1961) suggested the application of the Bonferroni inequality to multiple comparisons of means. To apply this method in its simplest form, we need only decide at what level we wish to control α_E and then set the nominal α_C for each pairwise comparison equal to $\alpha_E/[k(k-1)/2]$. (Fisher, 1935, also suggested this approach.) If, in our example, we do not want the experimentwise error rate to exceed .05, we set the nominal α_C equal to .05/6 = .0083. The easiest way to achieve this is to refer to a table of the Bonferroni t statistic (see Kirk, 1982 or Miller, 1981). For $\alpha_E = .05$, $C = k(k-1)/2 = 6$ comparisons, and $N - k = 36$ degrees of freedom, we find by linear interpolation that the critical value for a two-sided test is $t^B_{C,N-k;1-\alpha_E} = t^B_{6,36;.95} = 2.80$. Therefore, as was the case with the Scheffé and Tukey procedures, only the comparisons of the G and C groups and of the P and C groups are statistically significant.

For significant comparisons, simultaneous $100(1 - \alpha_E)\%$ confidence intervals of the form shown in Equation 7 can be constructed, with $t^B_{C,N-k+1-\alpha_E}$ substituted for $S_{k-1,N-k+1-\alpha_E}$. In the case of the Bonferroni approach, each interval could also be interpreted as an individual (non-simultaneous) $100(1 - \alpha_C)\%$ confidence interval for the mean difference in question.

The Bonferroni approach is extremely flexible. It can be applied to cases in which the researcher wishes to use an unequal allocation of error rates (i.e., a different value of α_C for each contrast) or to perform one-sided tests. For these more complicated applications, the best table of critical values is that of Dayton and Schafer (1973). The Bonferroni approach is not limited to pairwise comparisons, but can be applied to any contrasts of interest. Because this MCP controls α_E at a nominal value, there is no reason to precede Bonferroni t-tests with an F-test.

Fisher's Protected t-tests. The protected t-test procedure, also called the Least Significant Difference (LSD) test, is unlike the procedures described above in that it is a sequential or stagewise procedure. First, an F-test is performed at the desired α_E level, say, .05. If it is found to be significant, all pairwise t-tests are performed, each with $\alpha_C = \alpha_E = .05$. The determination of whether the t-test are to be conducted depends on the results of the prior F-test. It is not permissible to omit the F-test here, as is allowed in the Scheffé approach. (Note that the term "Least Significant Difference Test" is sometimes applied to multiple t-tests performed without a prior F-test as well.)

In the present example, the F-test was found to be significant. Therefore, in accordance with the protected t approach, all six pairwise t-tests are to be computed and compared to $t_{N-k;1-\alpha_E/2} = t_{36;.975} = 2.03$. We find that the P - G, G - I, and G - C comparisons are statistically significant.

Because the protected t-test procedure tends to lead to a larger number of statistically significant comparisons than many of its competitors, it has sometimes been recommended as a powerful MCP (e.g., Carmer & Swanson, 1973; Cohen & Cohen, 1975). In fact, its apparent power is, at least in part, a result of poor Type I error control: although this MCP provides better Type I error control than multiple t-tests without a prior F, use of the protected t procedure can still lead to excessive Type I error rates. Contrary to what is often believed, the policy of performing pairwise t-tests only when the F-test is significant does not, in general, ensure that the experimentwise error rate will be controlled at α_E, the Type I error rate for the F-test. When there are more than k = 3 groups, the error rate will be controlled at α_E in the complete null case; that is, when all k means are identical in the population.

However, the true situation may be a partial null case: some pairs of population means may differ, whereas others do not. Suppose, for instance, that we were conducting an experiment with k = 5 groups and that the values of the five population means were as follows: 10, 3, 3, 3, 3. If we found the F-test significant, this would not be a Type I error. We would then perform k(k - 1)/2 = 10 pairwise t-tests, each at $\alpha_C = .05$. In

doing so, we would have the opportunity to make Type I errors by falsely concluding that the identical means were different. In fact, the number of Type I errors could be as large as $(k - 1)(k - 2)/2 = 6$, the number of distinct pairwise comparisons among the $k - 1 = 4$ means with population values of 3. The occurrence of these second-stage Type I errors leads to an inflated experimentwise error rate for this procedure (see Ryan, 1959, 1980; Zwick & Marascuilo, 1984).[4] It should be mentioned, however, that despite the liberalized error control in the second stage, a significant F does not imply the existence of a signficiant pairwise comparison (see Games, 1971, p. 558).

Hayter (1986) derived an exact expression for the maximum experimentwise error rate that can be attained for Fisher's protected t procedure in the equal-n case. (The same quantity serves as an upper bound for the unequal-n case). With a nominal α_E of .05 and infinite degrees of freedom, the maximum experimentwise error rate is found to be .1222 for $k = 4$, .5715 for $k = 10$, and .9044 for $k = 20$. Some empirical evidence on the experimentwise error rates of the protected t-test procedure in partial null cases is provided by Carmer and Swanson (1973). Computer simulation techniques were used to estimate the experimentwise error rates of the method in 14 partial null configurations said to be "somewhat representative of situations found in actual experiments in the agricultural sciences" (p. 69). The number of means was 5, 10, or 20. With the nominal α_E set at .05, the $\hat{\alpha}_E$ values ranged from .023 to .455. Five of the 14 error rates were greater than .15. Thus, it can be

demonstrated both theoretically and empirically that (for k > 3) the protected t method, unlike the Scheffé and Tukey procedures, is not assured to control the error rate in partial null cases.

Because this MCP is a sequential procedure, involving different levels of error control at each stage, it is impossible to derive confidence intervals corresponding to the protected t procedure. The unavailability of confidence intervals is a property of all MCPs in which the performance of certain comparisons is concingent on the significance of other comparisons or of an omnibus test, such as the F-test.

Newman-Keuls Test. Another commonly used sequential procedure is the Newman-Keuls test. Like the Tukey MCP, this method involves rank-ordering the means and performing the test shown in Equation 8 (or the equivalent test based on Equation 4) to compare the largest and smallest means. If this initial range test is significant, further comparisons are made using reduced critical values: the closer two means are to each other in the ranking, the less stringent the criterion for significance. For testing the range of p means, where $p \leq k$, the critical value is $q_{p,N-k;1-\alpha_E}/\sqrt{2}$, assuming here that test statistics are of the form shown in Equation 4. That is, in determining the critical values for all tests that follow the first one, we simply ignore the fact that the experiment has k groups and use the same critical value we would use if we were performing Tukey's test with p groups.

For example, with four groups of 10 subjects as in the present study, the first step in performing the Newman-Keuls test is to rank-order the means from smallest to largest ($\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, $\bar{X}_4$). If we set the nominal α_E equal to .05, $\bar{X}_1$ is compared to $\bar{X}_4$ using a critical value of $q_{k,N-k;1-\alpha_E}/\sqrt{2} = q_{4,36:.95}/\sqrt{2} = 3.81/\sqrt{2} = 2.70$. If this test is significant, the tests of $\bar{X}_2$ versus $\bar{X}_4$ ar' $\bar{X}_1$ versus $\bar{X}_3$ (ranges of p = 3 means) are performed using a critical value of $q_{p,N-k;1-\alpha_E}/\sqrt{2} = q_{3,36:.95}/\sqrt{2} = 3.46/\sqrt{2} = 2.45$. Finally if all these tests prove to be significant, we test $\bar{X}_1$ versus $\bar{X}_2$, $\bar{X}_2$ versus $\bar{X}_3$, and $\bar{X}_3$ versus $\bar{X}_4$ with a critical value of $q_{2,36:.95}/\sqrt{2} = 2.87/\sqrt{2} = 2.03$. If at any point, a range of p means is found to be nonsignificant, no comparisons of means within that range are performed. Thus, no range included in a nonsignificant range can be declared significant. In the case of unequal sample sizes, the Newman-Keuls MCP can be modified in the same manner as Tukey's test.

In the present example, the four means, from lowest to highest, are $\bar{X}_G = 25$, $\bar{X}_I = 36$, $\bar{X}_P = 40$, and $\bar{X}_C = 41$. The test of the range of all k = 4 means, shown in Equation 8, was statistically significant. We can therefore proceed to test the two ranges of p = 3 means with a critical value of 2.45, as described above. The P - G comparison is found to be significant, whereas the I - C comparison is not. The G - I comparison is then tested with a critical value of 2.03 and is found to be statistically significant. (The P - I and P - C comparisons are not tested because they fall within a nonsignificant range.)

As is the case with protected t-tests, the Newman-Keuls test is often mistakenly believed to be a powerful procedure, because it tends to produce a larger number of statistically significant differences than some of its competitors. However, because its error control becomes less stringent at each stage, the Newman-Keuls test, like the protected t procedure, does not maintain the experimentwise error rate at the nominal α_E for all possible configurations of true mean values (unless $k \leq 3$; see footnote 4). Some empirical evidence of its lack of error control in partial null cases is provided by Ramsey (1981) who found $\hat{\alpha}_E$ to range from about .13 to .15 for k = 6 and a nominal α_E of .05. Another popular stagewise MCP that is based on the Studentized range is Duncan's multiple range test. This procedure provides even less stringent error control than the Newman-Keuls test. For the same reason cited in connection with the protected t procedure, confidence intervals cannot be derived for the Newman-Keuls or Duncan tests.

Relation of MCPs to the ANOVA F-test. It is important at this point to summarize the relation between pairwise MCPs and the ANOVA F-test: If a researcher is interested only in pairwise comparisons between means, there is no need to perform an F-test. In fact, an F-test and an MCP may produce inconsistent results: The F-test may be significant when there are no significant pairwise comparisons and, except in the case of the Scheffé MCP, a pairwise comparison may be significant when the F is not significant. It is often believed that a prior F-test is necessary to

achieve adequate Type I error control. However, the Tukey and Bonferroni MCPs, which have been recommended here as the most desirable procedures, provide experimentwise error control without a prior test. Requiring a significant F prior to the performance of these tests will cause an unnecessary reduction of their Type I error rates and a corresponding loss in power. In the case of the Scheffé MCP, a prior F serves only as a labor-saving device, but does not affect the comparisons found significant. The protected t procedure does rely on a prior F to maintain the experimentwise error control at the nominal level in the complete null case. Similarly, a Studentized range test of the largest versus the smallest means must precede all other tests in the Newman-Keuls MCP. However, as stated above, these two sequential procedures should be avoided because, for $k > 3$, they provide inadequate error control in partial null cases <u>despite</u> the use of prior overall tests.

In practice, most MCPs are now conducted using statistical software packages which perform MCPs only in conjunction with an ANOVA F-test. However, the results of the F-test need not be used as a criterion for consideration of MCP results; rather, the researcher can proceed directly to the MCP results, regardless of whether the F is significant.

<u>Evaluation of the Five MCPs</u>. Five MCPs have been described for the case of k independent samples: a) Scheffé's procedure, b) Tukey's Studentized range test and the Tukey-Kramer modification, c) the Dunn-Bonferroni approach, d) Fisher's protected t-tests, and e) the Newman-Keuls test. A

summary of the properties of these methods is given in Table 3. How can we choose among these procedures?

Insert Table 3 about here

Ideally, we would like to select a method that provides a powerful test while maintaining adequate Type I error control, requires few statistical assumptions, and is easy to apply. All five procedures can be performed by hand, although the Newman-Keuls becomes unwieldy for large k, and all can be conducted using software packages such as SPSS^{x} (SPSS, Inc., 1986), SAS (SAS Inc., 1988), and BMDP (Dixon, Brown, Engelman, Hill, & Jennrich, 1988). They all require the assumption that the observations are independent random samples from normal populations with equal variances.

One way in which the five MCPs can be distinguished is in terms of the number of statistically significant comparisons. Two of the MCPs -- the protected t procedure and the Newman-Keuls -- yielded three significant comparisons for the psychotherapy data, whereas the remaining MCPs found only two comparisons to be significant. However, this evidence alone is not sufficient to draw conclusions about the relative power of the methods. The Type I error rates of the MCPs must also be considered. The protected t-tests and Newman-Keuls test can be ruled out as acceptable procedures because (for $k > 3$) they do not control the experimentwise Type I error rate at the nominal α_E for all possible configurations of true mean values. It should be stressed that there is no reason (other than tradition) that the Type I error rate need be controlled at .05. However, it is important

to choose a procedure that allows the researcher to control the error rate at some prespecified level. The protected t-test and Newman-Keuls procedures do not satisfy this criterion. In addition, the conditional nature of these MCPs makes the derivation of confidence intervals impossible.

The remaining three MCPs provide adequate Type I error control for pairwise contrasts. Because the test statistics for these MCPs are identical, we can compare their power for $k = 4$ groups, $N - k = 36$ degrees of freedom, and $\alpha_E = .05$ by comparing their critical values for the example. The Scheffé, Bonferroni, and Tukey critical values for test statistics of the form of Equation 3 or 4 were 2.93, 2.80, and 2.70, respectively, indicating that the Tukey method is the most powerful. For performing the set of all pairwise comparisons, the superiority in power of the Tukey (and Tukey-Kramer) methods to the Bonferroni and Scheffé methods holds in general (Miller, 1985; Stoline, 1981); the superiority of the Bonferroni to the Scheffé methods nearly always holds, with some exceptions occurring at small values of $N - k$. (For a fixed value of $N - k$, the discrepancies between the critical values increase with k. For instance, with $N - k = 36$ degrees of freedom as above and $k = 6$, the critical values for the Scheffé, Bonferroni, and Tukey methods are 3.52, 3.15, and 3.01, respectively; for $k = 10$, the values are 4.40, 3.55, and 3.37. For fixed values of k, the disparities between the critical values decrease slightly as the error degrees of freedom increase.) These three MCPs can also be applied to complex contrasts. In most practical, situations, the

Bonferroni method will have the highest power for tests of this kind; followed in order by the Scheffé and Tukey methods (Miller, 1981). (The LSD test could be extended to apply to complex contrasts, but this would compound its lack of Type I error control. Extension of the Newman-Keuls test to complex contrasts would not be straightforward.)

The Tukey or Tukey-Kramer approach is therefore recommended as the best method, in general, for performing pairwise comparisons in the normal equal-variance case. There are, however, special circumstances in which the Bonferroni MCP may be preferred. If only a subset of all pairwise comparisons is to be performed, the Bonferroni approach may be more powerful than the Tukey method. For example, if only three of the six pairwise comparisons in the psychotherapy study had been of interest, the Bonferroni critical value would have been $t^{B}_{C,N-k;1-\alpha_E} = t^{B}_{3,36;.95} = 2.52$, which is smaller than the Tukey critical value of 2.70. Furthermore, the Bonferroni approach, unlike the Tukey method, controls the Type I error rate in a comparisonwise fashion, which may be desirable if conclusions are to be based in the truth of individual statements. Also, the method can provide efficient tests of one-sided hypotheses and can accommodate unequal allocation of error rates, which may be useful in certain applications.

As a postscript to this evaluation of MCPs, it must be noted that there do exist stepwise MCPs that control the Type I error rate at a pre-specified α_E and are more powerful than the Tukey and Bonferroni methods (see Hochberg & Tamhane, 1987; Shaffer, 1986). However, application of these methods may require more effort than most researchers are willing to

invest. A more important drawback is that stepwise methods do not allow the construction of confidence intervals, which are extremely useful for the interpretation of results. For these reasons, single-step procedures are recommended here.

MCPs for Use Under Violation of the Equal-Variance and Normality Assumptions

The five MCPs described in the preceding sections are based on the assumptions of normality and equality of variances. If these assumptions are violated, neither the MCPs described above nor the ANOVA F-test are strictly valid. Alternative procedures that can be substituted in these cases are discussed in this section. In determining whether these alternative methods are required, it is important to consider that slight deviations from normality have been found to have little effect on the power and Type I error rates of normal theory ANOVA-based procedures, except when sample sizes are very small. On the other hand, relatively small departures from variance equality can have substantial effects, particularly when sample sizes are unequal.

MCPs for the Normal Unequal-Variance Case. A number of MCPs have been proposed for the normal unequal-variance case (see Dunnett, 1980; Games, Keselman, and Rogan, 1981, and Tamhane, 1979). Many of these are based in Welch's (1938) modification of the t-test, which requires that a test statistic of the form

[9] $$t^*_{ll'} = \frac{\bar{X}_l - \bar{X}_{l'}}{\sqrt{\frac{s_l^2}{n_l} + \frac{s_{l'}^2}{n_{l'}}}}$$

be compared to the desired percentile of the t distribution with $\nu_{ll'}$ degrees of freedom, where

[10] $$\nu_{ll'} = \frac{(s_l^2/n_l + s_{l'}^2/n_{l'})^2}{s_l^4/[n_l^2(n_l - 1)] + s_{l'}^4/[n_{l'}^2(n_{l'} - 1)]} .$$

Non-integer values of $\nu_{ll'}$ are rounded to the nearest integer. A simple way to apply this procedure to the case of multiple comparisons is to perform all k(k - 1)/2 tests of this kind, controlling α_E via the Bonferroni inequality (Ury and Wiggins, 1971). The procedure is somewhat cumbersome to perform by hand because the degrees of freedom, $\nu_{ll'}$, must be recomputed, and a new critical value, $t^B_{C,\nu_{ll'};1-\alpha_E}$, obtained for each comparison. In practice, however, the significance probabilities (p-values) for Welch t-tests can be obtained from packaged software, such as the SPSS^x T-TEST program (SPSS, Inc., 1986), the SAS TTEST program (SAS Institute, Inc., 1988), or BMDP7D (Dixon, et al., 1988). The Bonferroni inequality can then be applied by declaring significant those comparisons for which the p-value is less than α_E /[k(k - 1)/2].

For illustration, the procedure will be applied to the data of Table 2. For the G - C comparison,

$$t^*_{GC} = \frac{25 - 41}{\sqrt{\frac{95.33}{10} + \frac{94.22}{10}}} = -3.68$$

$$\text{and } \nu_{GC} \frac{(95.33^2/10 + 94.22^2/10)^2}{95.33^4/[10^2(9)] + 94.22^4/[10^2(9)]} = 17.97$$

The appropriate critical value is $t^B_{C,\nu_{ii'}:1-\alpha_E} = t^B_{6,18:.95} = 2.97$; therefore, the contrast is again found significant. The t* values for the P - G, P - I, P - C, G - I, and I - C comparisons are 3.40, .90, -.23, -2.50, and -1.14, respectively. The value of $\nu_{ii'}$ is, in each case, 18 when rounded to the nearest integer, leading, once again, to a critical value of 2.97. Therefore, only the G - C and P - G comparisons are statistically significant. An alternative to the Bonferroni approach is the Tukey-type MCP developed by Games & Howell (1976). However, the Type I error rates for this MCP sometimes exceed their nominal levels to a small degree (Dunnett, 1980; Tamhane, 1979). A Scheffé-type MCP has been developed for the unequal-variance case as well (Brown & Forsythe, 1974), but its power is low for pairwise comparisons. (Also, Rubin, 1983, has described some problems associated with the approximation proposed by Brown & Forsythe.) Although the methods mentioned here do not require a prior test, it should be noted that overall hypothesis tests analogous to F-test exist for the one-way ANOVA model in the unequal-variance case (e.g., Welch, 1951; see Rubin, 1983).

MCPs for the Nonnormal Case. In this section; nonparametric MCPs for the nonnormal case are described. All the procedures in this section

(Equations 11-14) are based on large-sample approximations. As a rule of thumb, it is suggested that they be used with caution for $N < 20$. If there is reason to believe that normality does not hold, one option is to use Scheffé-type MCPs based on the Kruskal-Wallis (1952) rank analogue to parametric ANOVA. This approach, which was presented by Nemenyi (1963) and is described in Marascuilo and McSweeney (1977) and Miller (1981), is illustrated for the data of Table 2. To perform a Kruskal-Wallis test, the observations must first be ranked from 1 to N, ignoring group membership. Midranks are assigned to ties. The ranked observations and the sums of the ranks for each group (R_i) are shown in Table 4.

Insert Table 4 about here

The Kruskal-Wallis statistic is computed as follows:

$$[11] \quad H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} R_i^2 - 3(N+1)$$

$$= \frac{12}{40(41)} [249.5^2/10 + 101.5^2/10 + 207.5^2/10 + 261.5^2/10] - 3(41)$$

$$= 11.63.$$

Because the value of H exceeds $\chi^2_{k-1;1-\alpha_E} = \chi^2_{3;.95} = 7.81$, the null hypothesis of no group differences is rejected at $\alpha_E = .05$. In order to compare groups i and i′, the following test statistic is computed:

$$[12] \quad Z_{ii'} = \frac{\bar{R}_i - \bar{R}_{i'}}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}},$$

where $\bar{R}_i$ and $\bar{R}_{i'}$ are the mean ranks for groups i and i′, respectively. The critical value is

[13] $$S' = \sqrt{\chi^2_{k-1;\,1-\alpha_E}} .$$

This is analogous to the use of S (Equation 6) as a critical value for parametric ANOVA. A comparison is statistically significant if $Z > S'$ or $Z < -S'$. The $Z_{ii'}$ statistic for comparing the G and C groups is

$$Z_{GC} = \frac{101.5 - 261.5}{\sqrt{\frac{40(41)}{12}\left(\frac{2}{10}\right)}} = -3.06,$$

which exceeds $S' = \sqrt{7.81} = 2.79$. Therefore, as in the previous analyses, it is concluded that the combined therapy and group therapy groups differ in ability to function. The $Z_{ii'}$ values for the P - G, P - I, P - C, G - I, and I - C comparisons are 2.83, .80, -.23, -2.03, and -1.03, respectively. Again, only the G - C and P - G comparisons are statistically significant. For a more precise test, a correction for ties should be used in the computation of both the H statistic and the Z-statistics (see Marascuilo & McSweeney, 1977, pp. 302, 318). Use of the correction increases the likelihood of rejecting the null hypothesis. In the present example, the use of the correction would not have changed the conclusions.

As was true of the Scheffé approach in the parametric case, it is possible to find that the Kruskal-Wallis test is significant, but that no pairwise comparisons are significant. Another similarity to the parametric case is that no additional risk of Type I error is incurred if the

Kruskal-Wallis test itself is omitted; the researcher can proceed directly to the performance of pairwise MCPs using S´ as a critical value. However, if only pairwise comparisons are of interest, a more powerful test can be achieved by employing the Bonferroni critical value, $t^{B}_{C,\infty:1-\alpha_E}$, where "$\infty$" indicates that the critical value for infinite degrees of freedom should be used (Dunn, 1964). The Bonferroni critical value for the example is $t^{B}_{6,\infty:.95}$ = 2.64, as compared to S´ = 2.79. A still more powerful approach is the joint-ranking analog to Tukey's test, for which the critical value is $q_{k,\infty:1-\alpha_E}/\sqrt{2} = q_{4\infty:.95}/\sqrt{2} = 3.63 - 2.57$. This method was proposed for the equal-n case by Nemenyi (1963, see Levy, 1979; Miller, 1981) but provides a good approximation in the case of unequal sample sizes (Miller, 1985; Zwick & Marascuilo, 1984).

The three MCPs described above make use of ranks based on all k groups (joint ranking). This may be considered undesirable because it leads to a situation in which the hypothesis test for each pair of populations is conditional on the location of the other k - 2 populations in the study. (In addition, Oude Voshaar [1980] has shown that, because of this property, the experimentwise Type I error rate for the rank analog to Tukey's test can exceed the nominal α_E in partial null cases.) A related disadvantage of MCPs based on joint ranking is that it is nearly impossible to obtain confidence intervals in the original metric of the observations. The complexity of the calculations is a result of the dependence of each pairwise comparison on all observations in the study (Miller, 1981, pp. 168-169). Because of these properties, the researcher may prefer an

MCP in which a separate ranking is performed for each comparison (pairwise ranking). For example, all pairwise independent-sample Wilcoxon tests could be performed, controlling α_E via the Bonferroni inequality (see Dunn, 1964). A Z-statistic of the form shown in Equation 12 is then compared to $t^B_{C,\infty:1-\alpha_E}$. (Although the Z-statistic in Equation 12 does not resemble a conventional Wilcoxon test, it is equivalent to the normal approximation to the Wilcoxon test when pairwise ranking is used.) The only difference between this MCP and the previously described rank-based Bonferroni approach is that, in this MCP, the ranks for a given comparison are based only on the groups included in that comparison. A more powerful test is the k-sample Steel-Dwass procedure, which is an analog to Tukey's test based on pairwise ranking (see Miller, 1981; Hochberg & Tamhane, 1987). The use of MCPs based on pairwise, rather than joint ranking, allows the construction of confidence intervals (Miller, 1981, pp. 145-146). It should be noted that joint and pairwise ranking procedures will not necessarily lead to the same conclusion (see Dunn, 1964, Hollander & Wolfe, 1973, and Hochberg & Tamhane, 1987 for further discussion of this issue).

Rank procedures are useful when there is reason to believe that normality does not hold. If normality is violated, rank tests can be substantially more powerful than parametric tests. (Under some circumstances, related nonparametric methods called normal score procedures are more powerful than rank tests; see Marascuilo and McSweeney, 1977.) Unfortunately, rank methods, like their parametric counterparts, do not provide adequate Type I error control if the equal-variance assumption is

not met (e.g., see Van der Vaart, 1961). They are therefore not well-suited to the situation in which both the normality and equal-variance assumptions are violated. A procedure that may perform adequately in this situation is the MCP based on the all pairwise median tests. The first step is to find the median for the combined data for each pair of groups. Then the observations in each pair of groups are categorized according to whether they fall above or below the median for that pair. For example, if the data for the G and C groups in Table 2 are combined and ranked, the median is found to be 32.5. Two observations in the G group and eight observations in the C group are found to be above the median. For equal sample sizes, the appropriate test statistic is

$$[14] \qquad Z'_{ii'} = \frac{\hat{p}_i - \hat{p}_{i'}}{\sqrt{1/2n}} ,$$

where $\hat{p}_i$ and $\hat{p}_{i'}$ represent the proportion of cases in groups i and i' that are above the median. These $Z'_{ii'}$ values are compared to $q_{k,\infty:1-\alpha_E}/\sqrt{2}$. (Critical values based on the Scheffé or Bonferroni approach could be used but would lead to less powerful tests.)

The test statistic for the G - C comparison is

$$Z'_{GC} = \frac{.8 - .2}{\sqrt{1/20}} = 2.68.$$

Z'_{GC} exceeds $q_{k,\infty:.95}/\sqrt{2} = q_{4,\infty:.95}/\sqrt{2} = 2.57$, so it can be concluded once again that the G and C groups differ. Analogous computations show that the P - G comparison is also significant, but the remaining four comparisons are not. Further discussion of pairwise median test is given by Hochberg and Tamhane

(1987). Another median-based MCP that has been proposed (see Miller, 1981) is conducted by obtaining the combined median for all k groups and then calculating what proportion of each group is above it. This procedure is subject to the problems associated with all joint ranking procedures and can also lead to nonsensical conclusions in some circumstances (see Hochberg & Tamhane, p. 269).

As an alternative to the traditional nonparametric MCPs for the non-normal case discussed in this section, it may be possible to transform the data and then apply standard MCPs. Another alternative is the application of robust MCPs, which involve t-like statistics based on estimators other than the sample mean and variance (Dunnett, 1982). Robust methods may be a good choice when both the normality and equal-variance assumptions are thought to be violated.

One-Factor Repeated Measures Designs

The example of Table 2 involved three independent groups, each of which was exposed to a different condition. Another commonly used design involves a single group of subjects examined under k different conditions or on k occasions. An analysis of variance can be performed to test the hypothesis that all k means are equal in the population. If an overall test of this kind is desired, the researcher must choose between two general analysis strategies: the multivariate approach and the univariate mixed-model approach. Useful discussions of the computational details of these analyses and of the issues involved in choosing between the two

approaches are given by in the chapter by Lewis in this volume and by Barcikowski and Robey (1984), Finn and Mattson (1978), McCall and Appelbaum (1973), and Vitaliano (1981). Only a brief description of the two methods is given here.

In the multivariate approach, the k original variables (one for each occasion or condition) are transformed to k - 1 new variables, each of which may represent a contrast of interest (see Morrison, 1976, pp. 145-146 for details). One possibility is to transform the k observations for each subject (x_1, x_2,....x_k) to differences between successive observations ($x_1 - x_2$, $x_2 - x_3$,...$x_{k-1} - x_k$). These k - 1 difference scores are then treated as a single multivariate observation and a one-sample Hotelling's (1931) T^2 is applied. In the univariate mixed-model approach, the analysis is treated as a Subjects x Conditions ANOVA. The appropriate F statistic is the mean square for subjects divided by the mean square for the Subjects x Conditions interaction.

Both the multivariate and mixed-model approaches require the assumption that the k observations for each subject are drawn from a multivariate normal distribution (Rouanet & Lépine, 1970) and that subjects are independently sampled. A disadvantage of the mixed-model approach is that, in order for the analysis to be valid, the variance-covariance matrix of the repeated measures must satisfy a condition called sphericity or circularity. This property is equivalent to equality of the variances of difference scores for all possible pairs of the k conditions included in the experiment (Huynh & Feldt, 1970). The multivariate approach does not

require this assumption, but will often be less powerful than the mixed-model analysis.

Fortunately, if a researcher is interested only in comparing pairs of conditions, it is not necessary to choose between the two analysis strategies or to be concerned about sphericity. The researcher need only perform all $k(k - 1)/2$ correlated-sample t-tests (or any subset of these), using an error term based only on the two groups being compared, and controlling α_E via the Bonferroni equality (Myers, 1979). To illustrate this approach, assume that the data of Table 2 represent a series of four measurements (which, for ease of reference, will continue to be denoted as P, G, I, and C) on a single group of $n = 10$ subjects. A correlated-sample t-test comparing conditions i and i′ can be calculated according to the following formula

[15] $$t_{ii'} = \frac{\bar{D}_{ii'}}{\sqrt{\frac{s^2_{D_{ii'}}}{n}}}$$

where $\bar{D}_{ii'} = \bar{X}_i - \bar{X}_{i'}$ is the mean difference between the two conditions and $s^2_{D_{ii'}}$ the variance of the difference scores $(x_i - x_{i'})$. The appropriate critical value of $t^B_{C,n-1;1-\alpha_E}$ can be obtained from a table of the Bonferroni t statistic. It can be shown that

[16] $$s^2_{D_{ii'}} = s^2_i + s^2_{i'} - 2r_{ii'}s_i s_{i'} ,$$

where $r_{ii'}$ is the correlation between the two sets of observations. However, the simplest way to calculate $s^2_{D_{ii'}}$ is to actually compute the difference scores and then calculate their variance. Table 5 shows all three sets of difference scores, along with their means and variances.

Insert Table 5 about here

A pairwise comparison of the G and C conditions can be conducted as follows:

$$t_{GC} = \frac{\bar{D}_{GC}}{\sqrt{\frac{s^2_{D_{GC}}}{n}}} = \frac{-16}{\sqrt{\frac{1.15}{10}}} = -47.18$$

For $n - 1 = 9$ degrees of freedom, $C = 6$ comparisons, and $\alpha_E = .05$, the critical value can be found by linear interpolation to be approximately 3.40. Therefore, we would conclude that there is a difference between the G and C conditions. The t-statistics for the remaining five pairwise comparisons, given at the bottom of Table 5, are also statistically significant. Confidence intervals could be constructed for these mean differences as well. This example demonstrates that the inadvertent application of an MCP intended for independent samples to a repeated-measures design can lead to a substantial reduction in power: The Bonferroni approach led to two significant comparison in the independent-sample case and six in the repeated measures analysis.

Examination of Equation 16 reveals the reason for this: If an independent sample test is mistakenly applied, the square of the denominator of the test statistic will be

$$MSW\left(\frac{2}{n}\right) = \frac{(n-1)s_{j}^{2} + (n-1)s_{j'}^{2}}{2(n-1)}\left(\frac{2}{n}\right) = \frac{s_{j}^{2} + s_{j'}^{2}}{n}$$

instead of

$$\frac{s_{D_{jj'}}^{2}}{n} = \frac{s_{j}^{2} + s_{j'}^{2} - 2r_{jj'}s_{j}s_{j'}}{n}$$

The term $2r_{jj'}s_{j}s_{j'}$ will be positive whenever the correlation between the two sets of measurements is positive, which is the case in most applications. Therefore, by using an independent-sample MCP, the researcher is forfeiting the opportunity to subtract a positive term from the error estimate.

An important property of the correlated-sample t-tests described above is that, unlike the sets of t-tests conducted in the independent-sample, equal-variance case, they do not make use of a common error term. It is because there is no pooled error term that the sphericity assumption is not needed for these pairwise comparisons (see Boik, 1981).[5] Because the Bonferroni t-tests require no prior F-test, the best procedure to follow if only pairwise tests are of interest is to perform the t-tests only. It should be noted that, in the multivariate approach to repeated measures ANOVA, pairwise comparisons of condition means performed via the Roy-Bose (1953) method reduce to correlated-sample t-tests of the form shown in Equation 15. The critical value, however, is larger than that used in the Bonferroni approach, leading to a more conservative test.

Some empirical evidence on the performance of the Bonferroni approach with separate error terms is given by Maxwell (1980). He compared Tukey's Studentized range test (with a pooled error term based on the Subjects x Conditions interaction), two modifications of Tukey's test, both of which make use of separate error terms, the Roy-Bose method associated with the multivariate approach, and the Bonferroni method described above. For conditions in which sphericity held, both Tukey's test and the Bonferroni approach performed well. However, when sphericity was violated, the only procedure that provided adequate power while controlling the Type I error rate at the nominal level was the Bonferroni method.

When normality cannot be assumed, nonparametric procedures can be applied. For instance, multiple sign tests or multiple Wilcoxon signed-rank tests can be conducted, using the normal approximation (see Marascuilo & McSweeney, 1977) and controlling α_E via the Bonferroni inequality. Alternatively, the multiple comparison approach associated with the Friedman (1937) model can be applied (Levy, 1979; Marascuilo & McSweeney, 1977; Miller, 1981).

Summary

In order to select the appropriate pairwise MCP for use in a one-factor ANOVA model, the researcher should have a good understanding of experimentwise and comparisonwise Type I error rates. When testing

multiple hypotheses, the experimentwise error rate is usually of primary interest, although there are occasions in which comparisonwise control is useful. These two methods of assessing error rates are related but not interchangeable.

Another important issue is the relationship between Type I error rate and power. The power of competing MCPs can not be compared meaningfully unless the MCPs have equivalent experimentwise Type I error rates. Therefore, the selection of MCPs solely on the basis of the number of statistically significant results they produce is not well-founded. If a researcher wants to increase power, he should not attempt to do so by allowing the Type I error to exceed the desired level, but by increasing sample size, increasing the homogeneity of the sample, and improving the quality of measurement (see Cohen, 1982). There is, however, no reason that the experimentwise Type I error rate need be set to .05. A larger error rate certainly may be acceptable in some situations. What is important is that the researcher know and report the level at which the error has been controlled.

For the case of independent samples drawn from normal populations with equal variances, the Tukey method and, for unequal sample sizes, the Tukey-Kramer modification, were recommended as the best procedures in most situations. For certain specialized applications, such as those requiring one-sided tests or unequal allocation of error rates, the Bonferroni method may be preferred. For the normal unequal-variance case, Welch t-tests were recommended, with the Bonferroni approach used to control the experimentwise error rate. For the nonnormal case, a number of rank

procedures were discussed. The performance of all pairwise Wilcoxon tests, with the experimentwise error rate controlled via the Bonferroni or Steel-Dwass approach, has certain advantages over the procedures based on joint ranking. For the case in which both normality and equality of variances are thought to be violated, the MCP based on all pairwise median tests may be a good choice. For one-factor repeated measures designs, dependent t-tests with separate error terms were recommended, with error control achieved through the Bonferroni approach. Nonparametric MCPs for this design include multiple sign tests, multiple Wilcoxon signed-rank tests, and MCPs based on the Friedman model. In all MCP applications, the computation of confidence intervals can provide a useful supplement to significance testing. For this reason, as well as simplicity of computation, single-step MCPs, rather than more powerful stepwise methods, were recommended.

This paper was limited to the discussion of pairwise MCPs in one-factor designs. MCPs for complex contrasts, two-way ANOVA designs, and special situations, such as comparing experimental groups to a control group, are discussed by Hochberg and Tamhane (1987) and Miller (1981). Information about Bayesian, decision-theoretic, and robust MCPs, as well as extensive discussion of stepwise MCPs, is given in Hochberg and Tamhane (1987).

References

Barcikowski, R. S. and Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. The American Statistician, 38, 148-150.

Berhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F-test of ANOVA. Biometrics, 31, 229-232.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. Psychometrika, 46, 241-255.

Brown, M. B. and Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics, 30, 719-724.

Carmer, S. G., and Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 68, 66-74.

Cohen J., and Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1975

Cohen, P. (1982). To be or not to be: Control and balancing of Type I and Type II error. Evaluation and Program Planning, 5, 247-253.

Dayton, C. M., and Schafer, W. D. (1973). Extended tables of t and chi square for Bonferroni tests with unequal error allocation. Journal of the American Statistical Association, 68, 78-83.

Dixon, W. J., Brown, M. B., Engelman, L., Hill, M. A., & Jennrich, R. I. (1988) (eds.). BMDP statistical software manual. Berkeley, CA: University of California Press.

Duncan, D. B. (1955). Multiple range and multiple F-tests. Biometrics, 11, 1-42.

Dunn, O. J. (1958). Estimation of the means of dependent variables. Annals of Mathematical Statistics, 29, 1095-1111.

Dunn, O. J. (1959). Confidence intervals for the means of dependent, normally distributed variables. Journal of the American Statistical Association, 54, 613-621.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6, 241-252.

Dunn, O. J. (1974). On multiple tests and confidence intervals. Communications in Statistics, 3, 101-103.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 75, 796-800.

Dunnett, C. W. (1982). Robust multiple comparisons. Communications in Statistics: Theory and Methods, 22, 2611-2629.

Einot, I., and Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. Journal of the American Statistical Association, 70, 574-583.

Finn, J. D., and Mattson, I. (1978). Multivariate analysis in educational research: Applications of the MULTIVARIANCE program. Chicago: National Educational Resources.

Fisher, R. A. (1935). The design of experiments (1st ed.). Edinburgh: Oliver and Boyd.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32, 675-701.

Gabriel, K. R. (1978). Comment: Multiple comparison power. Journal of the American Statistical Association, 73, 485-487.

Games, P. A. (1971). Multiple comparisons of means. American Educational Research Journal, 8, 531-565.

Games, P. A., and Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1, 113-125.

Games, P. A., Keselman, H. J. and Rogan, J. C. (1981). Simultaneous pairwise multiple comparison procedures for means when sample sizes are unequal. Psychological Bulletin, 90, 594-598.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. Annals of Statistics, 12, 61-75.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's Least Significant Difference test. Journal of the American Statistical Association, 81, 1000-1004.

Hochberg, Y., and Tamhane, A. (1987). Multiple comparison procedures. New York: Wiley

Hollander, M., and Wolfe, D. A. (1973). Nonparametric statistical methods. New York: John Wiley and Sons.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Huynh, H. and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. Journal of the American Statistical Association, 65, 1582-1589.

Keuls, M. (1952). The use of the 'Studentized range' in connection with an analysis of variance. Euphytica, 1, 112-122.

Kirk, R. E. (1982). Experiment design (2nd ed.). Monterey, CA: Brooks/Cole.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. Biometrics, 12, 307-310.

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47, 583-621.

Levy, K. J. (1979). Nonparametric large-sample pairwise comparisons. Psychological Bulletin, 86, 371-375.

Marascuilo, L. A., and McSweeney, M. (1977). Nonparametric and distribution free methods for the social sciences. Monterey, CA: Brooks/Cole. Maxwell, S.E. (1980). Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics, 5, 269-287.

McCall, R. B., and Appelbaum, M. I. (1973). Bias in the analysis of repeated measures designs: Some alternative approaches. Child Development, 44, 401-415.

Miller, R. G. (1981). Simultaneous statistical inference, (2nd ed.). New York: Springer-Verlag.

Miller, R. G. (1985). Multiple comparisons. In S. Kotz and N. L. Johnson (eds.), Encyclopedia of Statistical Sciences, Vol. 5, pp. 679-689. New York: Wiley.

Morrison, D. F. (1976). Multivariate statistical methods, (2nd ed.). New York: McGraw-Hill.

Myers, J. L. (1979). Fundamentals of Experimental Design, (3rd ed.). Boston: Allyn & Bacon.

Nemenyi, P. (1963). Distribution-free multiple comparisons. Unpublished doctoral dissertation, Princeton University.

Newman, D. (1939). The distribution of range in samples from the normal population, expressed in terms of an independent estimate of standard deviation. Biometrika, 31, 20-30.

Oude Voshaar, J. H. (1980). (k-1) - mean significance levels of nonparametric multiple comparisons procedures. The Annals of Statistics, 8, 75-86.

Ramsey, P. H. (1978). Comment: Multiple comparison power. Journal of the American Statistical Association, 73, 487.

Ramsey, P. H. (1981). Power of univariate pairwise multiple comparison procedures. Psychological Bulletin, 90, 352-366.

Rouanet, H., and Lepine, D. (1970). Comparison between treatments in repeated-measurement design: ANOVA and multivariate methods. British Journal of Mathematical and Statistical Psychology, 23, 147-163.

Rubin, A. S. (1983). The use of weighted contrasts in analysis of models with heterogeneity of variance. Proceedings of the Business and Economics Section of the American Statistical Association.

Roy, S. N. and Bose, R. C. (1953). Simultaneous confidence interval estimation. Annals of Mathematical Statistics, 24, 513-536.

Ryan, T. A. (1959). Multiple comparisons in psychological research. Psychological Bulletin, 56, 26-47.

Ryan, T. A. (1980). Comment on "Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic." Psychological Bulletin, 88, 354-355.

SAS Institute, Inc. (1988). SAS/STAT user's guide, release 6.03 edition. Cary, NC: Author

Scheffé, H. (1953) A method for judging all contrasts in the analysis of variance. Biometrika, 40, 87-104.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. Journal of Educational Statistics, 4, 14-23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

Smith, R. A. (1971). The effect of unequal group size on Tukey's HSD procedure. Psychometrika, 36, 31-34.

SPSS, Inc. (1986). SPSS^x user's guide (2nd ed.). Chicago: Author.

Stoline, M. R. (1981). The status of multiple comparisons: Simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. The American Statistician, 35, 134-141.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. Journal of the American Statistical Association, 74, 471-480.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University.

Ury, H. K., and Wiggins, A. D. (1971). Large sample and other multiple comparisons among means British Journal of Mathematical and Statistical Psychology, 24, 174-194.

Van der Vaart, H. R. (1961). On the robustness of Wilcoxon's two-sample test. In H. de Jonge (Ed.), Quantitative methods in pharmacology. Amsterdam: North-Holland, 1961.

Vitaliano, P. P. (1982). Parametric statistical analysis of repeated measures experiments. Psychoneuroendocrinology, 7, 3-13.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350-362.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Winer, B. J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

Zwick, R. and Marascuilo, L. A. (1984). Selection of pairwise comparison procedures for parametric and nonparametric analysis of variance models. Psychological Bulletin, 95, 148-155.

Footnotes

[1]According to the Bonferroni inequality, if $\alpha_c = .05/[k(k - 1)/2]$, then α_E cannot exceed .05; that is, $\alpha_E \leq k(k - 1)/2 \cdot \alpha_c = .05$. If α_c is small and the number of comparisons is not too large, the Bonferroni approach provides a surprisingly good upper bound, i.e., the bound does not exceed the true error rate by a large amount (Miller, 1981).

[2]All critical values for 36 degrees of freedom were obtained by linear interpolation between values for 30 and 40 degrees of freedom.

[3]"WSD" is sometimes used to refer to a different procedure developed by Tukey in which the critical values are obtained by averaging the critical values from the HSD and Newman-Keuls methods.

[4]For $k = 3$, the protected t and Newman-Keuls MCPs are assured to control α_E even in partial null cases; see Hayter, 1986; Hochberg and Tamhane, 1987; Shaffer, 1979; 1986. Also see Hochberg and Tamhane, 1987, p. 4 and elsewhere, for a detailed discussion of the power of the protected t method.

[5]Boik (1981) shows that if sphericity does not hold, use of a pooled error term will, in general, lead to unsatisfactory tests of individual contrasts even if one of the available corrections for nonsphericity is applied. Even under minimal departures from sphericity, these "corrected" tests can have poor power properties and Type I error rates that differ substantially from their nominal values.

Table 1

Hypothetical Data on 100 Simulated Experiments with k = 3

Number of significant comparisons (type 1 errors)	Number of experiments with the indicated number of significant comparisons	
	Multiple comparison procedure 1	Multiple comparison procedure 2
0	95	95
1	4	0
2	1	0
3	0	5
Total	100	100

Table 2

Hypothetical Data for Psychiatric Treatment Study*

	Pharmacologic Therapy (P)	Group Psychotherapy (G)	Individual Psychotherapy (I)	Combination of P and I (C)
	24	12	21	27
	29	13	25	28
	33	18	30	35
	36	20	32	37
	38	21	36	39
	42	29	37	44
	44	30	38	45
	47	30	42	46
	51	35	43	52
	56	42	56	57
$\bar{X}_i$	40	25	36	41
s_i^2	99.11	95.33	98.67	94.22
n_i	10	10	10	10

* $\bar{X}$ = 35.5, MSB = 536.07, MSW = 96.83, N = 40, t_{PG} = 3.41, t_{PI} = .91, t_{PC} = -.23, t_{GI} = -2.50, t_{GC} = -3.64, and t_{IC} = -1.14.

Table 3

Properties of Pairwise Multiple Comparison Procedures for Independent Samples in the Normal Equal-Variance Case

	Scheffé	Tukey HSD	Dunn-Bonferroni	Fisher's protected t-tests	Newman-Keuls
Allows control of experimentwise Type I error rate at preassigned level	Yes	Yes	Yes	Error rate is not controlled at nominal α_E for all possible configurationa of means	
Allows computation of confidence intervals	Yes	Yes	Yes	No	No
Allows efficient one-sided tests and unequal allocation of error rates	No	No	Yes	No	No
Can provide tests of complex contrasts	Yes	Yes	Yes	Yes, but Type I error control is inadequate	No
Power ranking for pairwise comparisons*	3	1	2	Not ranked because of inadequate Type I error control	
General critical value for t-statistics of Equation 4	$\sqrt{(k-1)F_{\alpha;k-1,N-k;1-\alpha_E}}$	$q_{k,N-k;1-\alpha_E}/\sqrt{2}$	$t^B_{C,N-k;1-\alpha_E}$	$t_{N-k;1-\alpha_E/2}$, t-tests performed only if $F > F_{\alpha;k-1,N-k;1-\alpha_E}$	$q_{p,N-k;1-\alpha_E}/\sqrt{2}$, $p \leq k$ No comparisons are performed within ranges declared non-significant
Type of table required	F	Studentized range	Bonferroni t	F, t	Studentized range
Critical value for example of Table 2 ($N = 40$, $k = 4$, $C = k(k-1)/2 = 6$, $\alpha_E = .05$)	2.93	2.70	2.80	2.03	2.70 for p = 4, 2.45 for p = 3, 2.03 for p = 2.

*This ranking holds in almost all cases. It is assumed that all k(k - 1)/2 pairwise comparisons are of interest.

Table 4

Ranked Data for Psychiatric Treatment Study[a]

	Pharmacologic Therapy (P)	Group Psychotherapy (G)	Individual Psychotherapy (I)	Combination of P and I (C)
	7	1	5.5	9
	11.5	2	8	10
	17	3	14	18.5
	20.5	4	16	22.5
	24.5	5.5	20.5	26
	28	11.5	22.5	31.5
	31.5	14	24.5	33
	35	14	28	34
	36	18.5	30	37
	38.5	28	38.5	40
R_i	249.5	101.5	207.5	261.5
n_i	10	10	10	10

[a] $Z_{PG} = 2.83$, $Z_{PI} = .80$, $Z_{PC} = -.23$, $Z_{GI} = -2.03$, $Z_{GC} = -3.06$, and $Z_{IC} = -1.03$.

Table 5

Difference Scores for the Data of Table 2 Treated as a Repeated Measures Design with n = 10

	P-G	P-I	P-C	G-I	G-C	I-C
	12	3	-3	-9	-15	-6
	16	4	-1	-12	-15	-3
	15	3	-2	-12	-17	-5
	16	4	-1	-12	-17	-5
	17	2	-1	-15	-18	-3
	13	5	-2	-8	-15	-7
	14	6	-1	-8	-15	-7
	17	5	-1	-12	-16	-4
	16	8	-1	-8	-17	-9
	14	0	-1	-14	-15	-1
$\bar{D}_{ii'}$	15	4	-1.40	-11	-16	-5
$s^2_{D_{ii'}}$	2.89	4.89	.70	6.67	1.15	2.36
$t_{ii'}$	27.90	5.72	-5.29	-13.47	-47.18	-10.29